anales de psicología, 2013, vol. 29, nº 2 (mayo), 301-310
http://dx.doi.org/10.6018/analesps.29.2.139031

ISSN edición impresa: 0212-9728. ISSN edición web (http://revistas.um.es/analesps): 1695-2294

Significación clínica: falsos positivos en la estimación del cambio individual

Antonio Pardo[1]* y Rodrigo Ferrer[2]

[1] *Universidad Autónoma de Madrid (Madrid, España)*
[2] *Universidad de Tarapacá (Arica, Chile)*

Resumen: Tanto en la investigación aplicada como en la práctica clínica es habitual tener que evaluar el cambio que experimentan los pacientes como consecuencia del tratamiento que reciben. En este trabajo se valora el comportamiento de varios métodos estadísticos diseñados para estimar ese cambio. La valoración se ha centrado en un aspecto al que todavía no se le ha prestado atención: la tasa de falsos positivos. Para ello, se ha simulado una situación de no-cambio (diseño pre-post sin cambio entre el pre y el post) y se ha valorado el comportamiento de nueve estadísticos distintos en ese escenario. Se han utilizado tres tamaños muestrales distintos (25, 50 y 100) y se han simulado 1000 muestras de cada tamaño. Para evaluar el comportamiento de los estadísticos elegidos se ha calculado el porcentaje de veces que cada estadístico ha detectado un cambio. Puesto que la situación simulada es de no-cambio, cualquier alerta de cambio debe ser considerada un falso positivo. Los resultados obtenidos son bastante llamativos: ninguno de los nueve estadísticos evaluados ofrece un comportamiento aceptable. Únicamente se consiguen resultados aceptables cuando se trabaja con la desviación típica de las diferencias pre-post y se aplican criterios clásicos en lugar de los propuestos por la literatura relacionada con la significación clínica.

Palabras clave: Cambio clínicamente significativo; diferencia mínimamente importante; falsos positivos.

Title: Clinical significance: false positives in the estimation of individual change.

Abstract: In applied research and in clinical practice we often need to assess the change experienced by patients as a result of the treatment they have received. This paper assesses the performance of several statistical methods designed to estimate such change. This study focuses on one aspect that still has not received attention: the rate of false positives. We have simulated a situation of no-change (pre-post design with no change between pre and post) in which the behavior of nine different statistics have been evaluated. Three different sample sizes (25, 50 and 100) were used and 1000 samples of each size were simulated. To evaluate the behavior of the chosen statistics we have calculated the percentage of times that each statistic has detected change. Since no-change is the simulated situation, any occurrence of change should be considered a false positive. Results are quite striking: none of the nine statistics evaluated offers an acceptable behavior. Good performance is achieved only when the standard deviation of pre-post differences and the traditional criteria are used and not when those proposed by the literature related to the clinical significance are used.

Key words: Clinically significant change; minimally important difference; false positives.

Introducción

Tanto en la investigación aplicada como en la práctica clínica es habitual tener que evaluar el cambio que experimentan los pacientes como consecuencia del tratamiento que reciben. La cuantificación de ese cambio posee una importancia crucial para poder estimar correctamente el efecto del tratamiento.

Los métodos tradicionalmente utilizados para valorar el efecto de un tratamiento (las pruebas de significación o contrastes de hipótesis y las medidas del tamaño del efecto) aportan información muy útil, pero no necesariamente informan sobre la importancia del efecto evaluado (Abelson, 1995; Cohen, 1994; Kirk, 1996). Las pruebas de significación permiten descartar el azar como fuente de explicación de los cambios observados, pero no permiten saber si ese cambio es o no clínicamente relevante porque la significación estadística depende parcialmente del tamaño muestral (Jacobson, Follette y Revensdorf, 1984; Thomson, 1993, 2002). Las medidas del tamaño del efecto intentan superar este inconveniente; de hecho, cuanto mayor es el efecto observado, más probable es que se corresponda con un cambio clínicamente relevante; pero dado que este tipo de medidas dependen de la variabilidad de las puntuaciones analizadas, un efecto grande no necesariamente se corresponde con un efecto importante (Jacobson, Roberts, Berns y McGlinchey, 1999; Kazdin, 1999, 2001).

Por otro lado, tanto las pruebas de significación como las medidas del tamaño del efecto se utilizan para analizar diferencias entre promedios grupales, no para identificar el cambio individual. Por tanto, este tipo de herramientas no permite saber si un sujeto concreto cambia o no, ni tampoco conocer el porcentaje de sujetos que cambian (Barlow, 1981; Jacobson y Truax, 1991).

Estas limitaciones en las herramientas estadísticas tradicionalmente utilizadas para valorar el efecto de una intervención han hecho que tanto el interés de los terapeutas como el de los investigadores aplicados se haya ido desplazando (aunque solo en parte y lentamente) desde la *significación estadística* hacia la *significación clínica* (ver Kazdin, 1977; Kendall, 1997, 1999; Ogles, Lunnen y Bonesteel, 2001).

Este interés por la *significación clínica* ha ido en aumento no solo en el ámbito psicológico, donde existe una larga tradición en el uso de cuestionarios para medir resultados, sino en el ámbito médico, donde el uso de cuestionarios para medir resultados (el interés por los resultados informados por el paciente) ha emergido con fuerza en las últimas dos décadas, muy particularmente en los estudios de calidad de vida y de satisfacción con el tratamiento (Crosby, Kolotkin y Williams, 2003; Fayers y Machin, 2000; Jaeschke, Singer y Guyatt, 1989).

En el ámbito psicológico, la *significación clínica* suele ir asociada al concepto de *cambio clínicamente significativo* (CCS) (Bergin, 1971; Jacobson et al., 1984; Kazdin, 1977; etc.). En este contexto, la significación clínica se refiere a "la importancia

*** Dirección para correspondencia [Correspondence address]:**
Antonio Pardo. Facultad de Psicología. Universidad Autónoma de Madrid. Cantoblanco. 28049 Madrid (España).
E-mail: antonio.pardo@uam.es

práctica del efecto de una intervención, es decir, a si una intervención produce alguna diferencia real en los clientes o en las personas que interactúan con él en su vida cotidiana" (Kazdin, 2001, p. 455). En el ámbito médico, la *significación clínica* suele ir asociada al término *diferencia mínimamente importante* (DMI) (Jaeschke et al., 1989; de Vet et al., 2010; Revicki, Hays, Cella y Sloan, 2008; Wyrwich, 2004; Wyrwich, Tierney y Wolinsky, 1999). En este contexto, la significación clínica se refiere a "la diferencia más pequeña entre las puntuaciones del dominio de interés que los pacientes perciben como beneficiosa y que podría aconsejar, en ausencia de efectos secundarios indeseables y coste excesivo, un cambio en el tratamiento del paciente" (Jaeschke et al., 1989, p. 408).

Tanto el CCS como la DMI están haciendo referencia a la variación mínima que debe darse en las respuestas de los sujetos para poder concluir que se ha producido un cambio clínicamente importante, relevante o significativo (Crosby, Kolotkin y Williams, 2003; Gatchel y Mayer, 2010; McGlinchey, Atkins y Jacobson, 2002; Turner et al., 2010). Puesto que los términos CCS y DMI son equivalentes, en adelante únicamente hablaremos de cambio clínicamente significativo (CCS).

Para estimar cuándo se produce un CCS se han utilizado diferentes estrategias (ver Bergin y Lambert, 1978; Ogles et al., 2001), pero en las últimas décadas ha surgido un creciente interés por realizar esa estimación a partir de la información recogida mediante escalas o cuestionarios (resultados informados por el paciente). Este interés se ha plasmado en la aparición de muchos y muy variados métodos diseñados con la intención de poder utilizar las *respuestas de los sujetos* para decidir cuándo se produce un CCS (ver Crosby et al., 2003; Turner et al., 2010).

Estos métodos pueden clasificarse en dos bloques: los basados en un *criterio externo al cuestionario* (*anchor-based*) y los basados en la *distribución de las propias puntuaciones del cuestionario* (*distribution-based*) (ver Crosby et al., 2003; Lydick y Epstein, 1993; Norman, Sridhar, Guyatt y Walter, 2001). Los primeros, los métodos que incluyen un referente externo (los llamaremos abreviadamente *externos*) intentan cuantificar el CCS a partir de la relación existente entre las puntuaciones del cuestionario y algún criterio clínico externo (por ejemplo, las valoraciones de expertos, o las valoraciones que los propios sujetos hacen en otros cuestionarios en los que expresan su propia percepción del cambio, o las puntuaciones de algún grupo de referencia, etc.). Los métodos sin referente externo (los llamaremos abreviadamente *internos*) intentan cuantificar el CCS sin otra información que las propias puntuaciones del cuestionario, aplicando estadísticos que permiten separar las variaciones relevantes o sistemáticas de las irrelevantes o atribuibles a las fluctuaciones aleatorias propias del azar muestral (Bauer, Lambert y Nielsen, 2004; Crosby et al., 2003; Jacobson y Truax, 1999).

A pesar de que en las comparaciones realizadas entre métodos *internos* y *externos* se observan ciertas coincidencias (Cella, Hahn y Dineen, 2002; Childs, Piva y Fritz, 2005; Crosby, Kolotkin y Williams, 2004; Guyatt, Osoba, Wu, Wyrwich y Norman, 2002; Norman, Sloan, y Wyrwich, 2003; Norman et al., 2001; Rejas, Pardo y Ruiz, 2008; Turner et al., 2010; Wyrwich et al., 1999), las discrepancias encontradas hacen pensar que se trata de métodos que están midiendo constructos distintos (Crosby et al., 2003; Kolotkin, Crosby y Williams, 2002). De hecho, no faltan quienes, basándose en los resultados de estas comparaciones, afirman que el CCS únicamente puede estimarse con métodos *externos*, pues solo ellos van acompañados de un criterio clínico que implica una definición de lo que se considera mínimamente importante. Desde este punto de vista, los métodos *internos* solo estarían indicando si el cambio es *estadísticamente fiable*, sin ningún tipo de referencia a su importancia clínica (Crosby et al., 2003; Turner et al., 2010; de Vet et al., 2006, 2007).

Esta es la razón que ha llevado a no pocos expertos a recomendar que el CCS sea estimado utilizando una *combinación* de métodos *internos* y *externos* (Cella et al., 2002; Crosby et al., 2003; de Vet et al., 2007; Jacobson y Truax, 1991; Kolotkin et al., 2002; Sheldrick, Kendall y Heimberg 2001). Por ejemplo, Jacobson y Truax (1991) proponen valorar el CCS en dos pasos: (1) estimando primero la cantidad de cambio mediante un índice de cambio fiable (método *interno* diseñado para valorar si el cambio observado supera el error de medida del cuestionario) y (2) estimando si el sujeto en cuestión ha pasado a estar más cerca de la media del grupo funcional que de la del grupo no funcional (método *externo*).

Al adoptar una estrategia *combinada* para estimar el valor del CCS se está asumiendo que los métodos *internos* permiten identificar cambios *estadísticamente fiables* (cambios dignos de ser tenidos en cuenta en una primera aproximación) y los métodos *externos* permiten decidir cuál de esos cambios fiables alcanza a ser clínicamente relevante. Para decidir que se ha producido un CCS deben cumplirse ambos criterios: el cambio debe ser estadísticamente fiable y clínicamente relevante.

En este escenario, lo razonable sería comenzar aplicando algún método *interno* para poder decidir, antes de aplicar más tarde un método *externo*, si un determinado cambio va más allá de las fluctuaciones aleatorias propias del azar muestral. Y lo que cabe esperar de un método de estas características es que ofrezca información lo más precisa posible, es decir, que ayude a separar correctamente los cambios estadísticamente fiables de los que no lo son.

En ese sentido, el objetivo de este estudio es ofrecer una valoración de los métodos *internos* disponibles para identificar cambios *estadísticamente fiables*. En concreto, nos proponemos evaluar el comportamiento de varios métodos *internos* en una situación de *no cambio* (en una situación en la que no existe efecto del tratamiento) con intención de establecer la tasa de falsos positivos asociada a cada uno de ellos, es decir, con intención de identificar la frecuencia con la que cada método indica que se ha producido un cambio estadísticamente fiable cuando en realidad no se ha producido tal cambio. Conocer esta información servirá para poder identificar métodos con excesiva tendencia a considerar como estadís-

ticamente fiables cambios que en realidad son solo producto del azar muestral.

Algunos de los métodos *internos* que nos proponemos evaluar ya han sido objeto de atención en varios estudios (ver, por ejemplo, Atkins, Bedics, McGlinchey y Beauchaine, 2005; Bauer et al., 2004; McGlinchey et al., 2002; Speer y Greembaum, 1995; Turner et al., 2010). Atkins et al., (2005), por ejemplo, han comparado entre sí cinco de estos métodos en situaciones simuladas de cambio entre el pre- y el post-tratamiento; y Turner et al., (2010) han comparado varios métodos *internos* con varios *externos* para establecer la correspondencia existente entre ellos en situaciones de cambio. Pero todos estos estudios se han limitado a valorar solo algunos de los métodos disponibles y, en lo que nos interesa aquí, siempre los han valorado en situaciones de *cambio entre el pre- y el post-tratamiento*. No hemos encontrado estudios que hayan evaluado el comportamiento de estos métodos en situaciones de no cambio y, consiguientemente, no hemos encontrado información sobre la tasa de falsos positivos que arroja cada uno de ellos. Conocer esa tasa es el principal objetivo de este trabajo.

Métodos *internos* para estimar el CCS

En nuestra revisión de los métodos *internos* disponibles para la estimación del CCS hemos encontrado nueve métodos[1]. Todos ellos han sido diseñados o pueden adaptarse para estimar el cambio individual en un diseño pre-post, que es el diseño que utilizaremos aquí para evaluarlos. Los dos primeros métodos representan una tipificación de las diferencias individuales pre-post; difieren entre ellos en la fuente de variabilidad que utilizan para realizar la tipificación. Los cinco siguientes (del 3 al 7) valoran los cambios tomando como referencia el error típico de medida. Los dos últimos (el 8 y el 9) se basan en los pronósticos de la regresión lineal. La Tabla 1 ofrece una descripción detallada de estos nueve métodos.

Tamaño del efecto individual (*TEI*)

Se trata de un estadístico inicialmente diseñado para valorar promedios que ha sido adaptado para valorar el cambio individual. El estadístico *TEI* valora el cambio individual dividiendo cada diferencia pre-post entre la desviación típica del pre-tratamiento. Cohen (1988) ha propuesto una especie de regla general que suele utilizarse como guía para valorar la diferencia entre dos promedios: valores en torno a .20 indican un efecto leve, en torno a .50 un efecto moderado y en torno a .80 un efecto grande. Al valorar el cambio *individual* se considera que los valores absolutos mayores que .20 representan un cambio clínicamente significativo (ver Crosby et al., 2003; Mathias et al., 2010), pero también se han utilizado otros criterios como .33 (Eton et al., 2004; Yost et al., 2005), y .6 y 1 (Wyrwich y Wolinsky, 2000).

Diferencia individual tipificada (DIT)

Este estadístico, también llamado *coeficiente de respuesta al tratamiento* o *índice de eficacia*, consiste en tipificar cada diferencia pre-post dividiéndola entre su desviación típica, es decir, entre la desviación típica de las diferencias. La interpretación de la magnitud del cambio se efectúa siguiendo la regla de Cohen (1988) para valorar el tamaño del efecto (Angst, Verra, Lehmann y Aeschlimann, 2008; Mathias et al., 2010; Norman, Stratford y Regehr, 1997; Stucki, Liang, Fossel y Katz, 1995).

El *estadístico de sensibilidad* (Guyatt, Bombardier y Tugwell, 1986) es idéntico a la *DIT* con la única excepción de que la desviación típica de las diferencias se obtiene a partir de una muestra de referencia (grupo control) en la que se asume que no existe cambio sistemático; en nuestro estudio, puesto que se parte de un escenario de no cambio, esto sería equivalente a calcular la desviación típica de las diferencias pre-post (igual que en la *DIT*); por tanto, todos los resultados relativos a la *DIT* son trasladables al *estadístico de sensibilidad*.

Diferencia tipificada de Wyrwich (*DTW*)

Este método se basa en la idea de que un cambio fiable debe ser mayor que el error típico de medida de la escala (*ETM*), el cual se obtiene a partir de la variabilidad y fiabilidad de las puntuaciones de la escala. En un diseño pre-post, Wyrwich y sus colaboradores recomiendan estimar la fiabilidad aplicando el coeficiente alfa de Cronbach a las puntuaciones de una muestra de referencia o, en su defecto, a las puntuaciones del pre-tratamiento (Wyrwich et al., 1999). Y para decidir cuándo se ha producido un cambio significativo el valor de la *DTW* se compara con los criterios ±1.00, ±1.96 y ±2.77 (Wyrwich, 2004). El criterio ±1 corresponde a un efecto de tamaño medio cuando la fiabilidad vale.75; el criterio ±1.96 corresponde a los valores asociados, en una curva normal tipificada, a un nivel de confianza del 95%; el valor ±2.77 se obtiene multiplicando 1.96 por la raíz cuadrada de 2 para compensar el hecho de que se están utilizando dos muestras (pre-post) en lugar de una.

Índice de cambio fiable (*ICF*)

Este método es, quizá, el más popular de todos los métodos propuestos para valorar la significación estadística de un cambio individual. También se basa en el error típico de medida, pero a diferencia del método *DTW*, incorpora tanto el error típico de medida del pre- como del post-tratamiento. Para valorar con este índice cuándo se produce un cambio significativo se han utilizado diferentes criterios, pero los más habituales han sido ±1.64 y ±1.96, que son los puntos de corte que corresponden en una curva normal a niveles de confianza del 90% y del 95%, respectivamente (Jacobson et al., 1984; Jacobson et al., 1999; Jacobson y Truax, 1991).

[1] Existen otros métodos aparte de los incluidos aquí. Hemos excluido un método basado en las *curvas de crecimiento* (Speer y Greenbaum, 1995) porque requiere más de dos mediciones (aquí nos centraremos únicamente en diseños pre-post). Y hemos excluido un método basado en el análisis de los residuos de la regresión lineal (Crawford y Howell, 1998) por tratarse de una aproximación descriptiva no comparable con el resto de estadísticos elegidos.

Tabla 1. Estadísticos evaluados y criterios de cambio estadísticamente fiable.

Estadístico	*Ecuación*	*Criterios de cambio*
1. Tamaño del efecto individual (*TEI*)	$TEI = \frac{D_i}{\sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}}}$	$\|TEI\| \geq .20$ $\|TEI\| \geq .50$ $\|TEI\| \geq .80$
2. Diferencia individual tipificada (*DIT*)	$DIT = \frac{D_i}{\sqrt{\frac{\Sigma(D_i - \bar{D})^2}{n-1}}}$	$\|DIT\|, \|ES\| \geq .20$ $\|DIT\|, \|ES\| \geq .50$ $\|DIT\|, \|ES\| \geq .80$
3. Diferencia tipificada de Wyrwich (*DTW*)	$DTW = \frac{D_i}{S_X\sqrt{1 - R_{XX}}}$	$\|D_i\| \geq 1.00$ $\|D_i\| \geq 1.96$ $\|D_i\| \geq 2.77$
4. Índice de cambio fiable (*ICF*)	$ICF = \frac{D_i}{\sqrt{ETM_X^2 + ETM_Y^2}}$	$\|ICF\| \geq 1.64$ $\|ICF\| \geq 1.96$
5. Gulliksen-Lord-Novik (*GLN*)	$GLN = \frac{(Y_i - \bar{X}_{ref}) - R_{XX(ref)}(X_i - \bar{X}_{ref})}{S_{X(ref)}\sqrt{1 - R^2_{XX(ref)}}}$	$\|GLN\| \geq 1.96$
6. Hageman-Arrindell (*HA*)	$HA = \frac{(Y_i - X_i)R_{DD} + (\bar{Y} - \bar{X})(1 - R_{DD})}{\sqrt{R_{DD}}\sqrt{2(MC_E)}}$	$\|HA\| \geq 1.65$
7. Edwars y Nunnally (*EN*)	$EN = [R_{XX}(X_i - \bar{X}) + \bar{X}] \pm 2S_X\sqrt{1 - R_{XX}}$	$Y_i \leq EN_{inf}$, $Y_i \geq EN_{sup}$
8. Intervalo para el pronóstico promedio (*ICPP*)	$ICPP = \hat{Y}_i \pm Z_{\frac{\zeta}{2}}\sqrt{MC_E\left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right]}$	$Y_i \leq ICPP_{inf}$, $Y_i \geq ICPP_{sup}$
9. Intervalo para el pronóstico individual (*ICPI*)	$ICPI = \hat{Y}_i \pm \left(t_{n-2;\frac{\zeta}{2}}\right)\sqrt{MC_E\left[1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right]}$	$Y_i \leq ICPI_{inf}$, $Y_i \geq ICPI_{sup}$

X_i = puntuación del sujeto *i* en el pre-test; Y_i = puntuación del sujeto *i* en el post-test; D_i = $X_i - Y_i$; *ref* = muestra de referencia sin cambio sistemático (equivale al pre-test); n = tamaño de la muestra; S = desviación típica; R_{XY} = coeficiente de correlación de Pearson entre X e Y; R_{XX} = coeficiente de fiabilidad alfa de Cronbach; *ETM* = error típico de medida; R_{DD} = fiabilidad de las diferencias D_i; MC_E = media cuadrática error (varianza de los residuos de la regresión lineal); $\hat{Y}_i$ = pronóstico de la regresión lineal de Y sobre X.

Método Gulliksen-Lord-Novick (*GLN*)

También se basa en el *ETM*, pero incorpora la media y la desviación típica de la población hipotética hacia la cual deberían tender los resultados en la medición post-tratamiento (es decir, incorpora el efecto de regresión a la media; Hsu, 1989, 1995, 1996). Si no se tiene información relevante sobre la media y la desviación típica de la población de referencia, se utiliza la media y la desviación típica del pre-tratamiento (Hsu, 1999). Se considera que ha ocurrido un cambio significativo cuando el valor absoluto de *GLN* es mayor que 1.96.

Método Hageman-Arrindell (*HA*)

La peculiaridad de este método respecto del anterior (tanto este método como el anterior, *GLN*, son modificaciones o correcciones del *ICF*) es que, en un intento por mejorar su precisión, incorpora a la ecuación la fiabilidad de las diferencias entre el pre- y el post-tratamiento y una corrección basada en la idea de riesgo máximo de Cronbach (probabilidad de que un sujeto sea clasificado incorrectamente). Dando a ese riesgo un valor inicial de .05 (calculado con el coeficiente de correlación fi), se llega a un punto de corte de 1.65 (Hageman y Arrindell, 1999). Es importante señalar que la aplicación de este método se encuentra condicionada a que la fiabilidad de las diferencias (R_{DD}) sea, como mínimo, de.40 (Hageman y Arrindell, 1999, p. 1173).

Método Edwards-Nunnally (*EN*)

Consiste en calcular el intervalo de confianza para la puntuación verdadera del pre-tratamiento tal como sugieren Edwards, Yarvis, Mueller, Zingale y Wagman (1978) o Nunnally (1975; Nunnally y Kotsch, 1983) y considerar que se ha

producido un cambio significativo cuando la puntuación del post-tratamiento se sale de ese intervalo de confianza (ver Speer, 1992).

Intervalo de confianza para el pronóstico promedio de la regresión lineal (*ICPP*)

Este método y el siguiente se basan en una lógica distinta a los anteriores. Aquí no se utilizan las diferencias individuales entre el pre- y el post-tratamiento, sino el pronóstico que se obtiene para el post al llevar a cabo un análisis de regresión del post sobre el pre. El análisis se lleva a cabo con una muestra de referencia en la que se asume que no ha habido cambio. Tras obtener el intervalo de confianza (el intervalo se calcula asumiendo que los pronósticos se distribuyen normalmente), se considera que ha habido un cambio significativo cuando la puntuación del post-tratamiento (Y_i) cae fuera de ese intervalo (Crawford y Garthwaite, 2006).

Intervalo de confianza para el pronóstico individual de la regresión lineal (*ICPI*)

Este método utiliza la misma lógica que el método *ICPP*, pero el intervalo de confianza se calcula para el pronóstico individual (el pronóstico individual y el pronóstico promedio son idénticos, pero el error típico del pronóstico individual es mayor que el del pronóstico promedio; ver, por ejemplo, Pardo y San Martín, 2010, p. 384) y no se basa en la distribución normal sino la distribución t con $n - 2$ grados de libertad (Crawford y Howell, 1998; Crawford y Garthwaite, 2006).

Procedimiento

Para evaluar el comportamiento de los nueve estadísticos elegidos hemos simulado un diseño pre-post o antes-después, es decir, un diseño en el que se realizan dos mediciones a los mismos sujetos: una primera antes de aplicar el tratamiento (medición pre) y una segunda después de aplicado el tratamiento (medición post). La peculiaridad de la situación simulada es que las puntuaciones se han generado asumiendo que entre los momentos pre y post no se ha producido más cambio que el atribuible a la variación aleatoria propia del azar muestral. Por tanto, el escenario simulado se corresponde con un diseño pre-post, sin grupo control y sin cambio sistemático entre momentos (es decir, sin efecto del tratamiento).

Para facilitar la comprensión de la simulación llevada a cabo hemos optado por utilizar una distribución que resultara familiar, en concreto, hemos elegido una distribución con las características del cociente intelectual, es decir, una distribución normal con media 100 y desviación típica 15. La simulación se ha realizado con el programa MATLAB, versión 2009b.

Para cada caso simulado se ha generado, en primer lugar, la puntuación del momento pre (X). Esta puntuación se ha obtenido sumando diez ítems generados a partir de una distribución normal multivariante imponiendo una correlación de .47 entre cada par de ítems, lo que se corresponde con un coeficiente alfa de Cronbach (R_{XX}) aproximado de .90 (similar a los niveles de consistencia interna recomendados para los cuestionarios o escalas utilizados en el ámbito clínico; ver, por ejemplo, Abad, Olea, Ponsoda y García, 2011, p. 114; Lambert y Ogles, 2004). A continuación se ha generado la puntuación del post[2] (Y) mediante una combinación de la variable X y una variable aleatoria complementaria con distribución normal. Las puntuaciones del momento post (Y) se han generado asumiendo una correlación (R_{XY}) de .80 con las del momento pre (X) (valor al que se aproxima la correlación test-retest entre las subescalas del WISC-IV; ver, por ejemplo, Williams, Weiss y Rolfhus, 2003).

Se han simulado muestras de tres tamaños distintos (25, 50 y 100) intentando representar los tamaños muestrales habitualmente utilizados en estudios de este tipo (ver, por ejemplo, Crawford y Howell, 1998). De cada tamaño muestral se han simulado 1000 muestras (3000 en total).

En cada replica se han realizado los cálculos necesarios para obtener los nueve estadísticos elegidos. En un escenario como el propuesto, en el que se asume que no existe efecto del tratamiento (escenario de no cambio), un estadístico diseñado para detectar cuándo se produce un cambio significativo o fiable debería llevar a la conclusión de que no se ha producido ningún cambio, salvo los atribuibles a las variaciones propias de azar muestral. Una tasa de falsos positivos mayor de la esperable por azar estaría indicando que el correspondiente estadístico tiende, más de lo que debería, a identificar como significativos cambios que en realidad son solo variaciones aleatorias. Por tanto, el siguiente paso del proceso ha consistido en calcular el número de veces que cada estadístico ha llevado a la conclusión de que se había producido un cambio significativo (falso positivo).

Finalmente se ha valorado el comportamiento de los diferentes métodos, es decir el comportamiento de cada estadístico con su correspondiente criterio, registrando la tasa de falsos positivos. Esta tasa se ha calculado a partir del valor absoluto de los cambios, es decir, considerando que se había producido un cambio tanto cuando la puntuación del post era mayor que la del pre como cuando era menor.

Los datos se han analizado con la versión 19 del programa informático IBM SPSS Statistics.

Resultados y discusión

La Tabla 2 ofrece un resumen descriptivo de los datos simulados. La tabla incluye, para cada tamaño muestral, las medias y desviaciones típicas de las puntuaciones del pre (X), de las del post (Y) y de las diferencias pre-post (D). La tabla también incluye el coeficiente de correlación de Pearson en-

[2] No se han generado ítems para obtener la puntuación del post-test; se asume que la consistencia interna es una propiedad relativamente estable de aplicación a aplicación.

tre las puntuaciones del pre y del post y la fiabilidad de las puntuaciones del pre (coeficiente alfa de Cronbach). Los resultados de esta tabla permiten constatar que los datos simulados se encuentran muy próximos a los valores de referencia utilizados para realizar la simulación.

Tabla 2. Medias (desviaciones típicas de las medias) de los datos simulados (*X* = pre; *Y* = post; *D* = diferencia pre-post).

Descriptivos	*n* = 25	*n* = 50	*n* = 100
$\overline{X}$	99.88(3.05)	99.98(2.10)	100.04(1.58)
$\overline{Y}$	99.86(3.04)	100.04(2.10)	100.04(1.58)
$\overline{D}$	.019(1.87)	-.065(1.37)	-0.001(.94)
Desv típ de *X*	14.80(2.09)	14.90(1.49)	14.96(1.05)
Desv típ de *Y*	14.80(2.18)	14.97(1.51)	14.97(1.05)
Desv típ de *D*	9.45(1.34)	9.40(.95)	9.45(.69)
Correlación *XY*	.791(.07)	.798(.05)	.799(.03)
Alfa de Cronbach en *X*	.890(.03)	.896(.02)	.897(.01)

La Tabla 3 muestra las medias y las desviaciones típicas de los nueve estadísticos evaluados. En el caso de los seis primeros, los valores de la tabla reflejan el resultado de aplicar las ecuaciones propuestas en la Tabla 1. En el caso de los estadísticos *EN*, *ICPP* e *ICPI*, los resultados de la tabla se refieren a los límites inferior y superior del correspondiente intervalo de confianza.

Tabla 3. Media (desviación típica) de los estadísticos evaluados.

Estadísticos	*n* = 25	*n* = 50	*n* = 100
1. *TEI*	.001(.64)	-.004(.63)	.000(.63)
2. *DIT*	.002(1.00)	-.007(1.00)	-.000(1.00)
3. *DTW*	.004(2.00)	-.014(1.99)	-.001(1.99)
4. *ICF*	.004(1.41)	-.010(1.41)	-.001(1.40)
5. *GLN*	-.003(1.40)	.010(1.38)	.000(1.39)
6. *HA*	-.008(.98)	.016(.97)	-.000(.98)
7. EN_{inf}	90.41(13.25)	90.52(13.46)	90.55(13.44)
EN_{sup}	109.36(13.25)	109.44(13.46)	109.52(13.44)
8. $ICPP_{inf}$ (95%)	95.19(11.88)	96.73(11.98)	97.67(12.00)
$ICPP_{sup}$ (95%)	104.53(11.87)	103.36(11.98)	102.41(11.99)
9. $ICPI_{inf}$ (95%)	81.11(11.81)	81.99(11.95)	82.16(11.98)
$ICPI_{sup}$ (95%)	118.63(11.80)	118.10(11.94)	117.91(11.98)

Tanto la medias como las desviaciones típicas que recoge la tabla se han obtenido promediando los valores de las 1000 muestras generadas con cada tamaño muestral. Por ejemplo, el primer valor de la tabla (*TEI* = .001) se ha obtenido promediando primero los 25 valores *TEI* correspondientes a cada una de las 1000 muestras simuladas (se obtienen así 1000 medias) y promediando a continuación esas 1000 medias iniciales; el valor de la correspondiente desviación típica (.64) es el resultado de promediar las desviaciones típicas de los *TEI* obtenidos en cada una de las 1000 muestras de tamaño 25.

Finalmente, la Tabla 4 ofrece, para cada estadístico evaluado, el porcentaje medio de falsos positivos. Estos porcentajes se han obtenido calculando el número de falsos positivos en cada muestra y promediando después el resultado de las 1000 muestras.

Tabla 4. Porcentaje medio (desviación típica) de falsos positivos.

Criterios de cambio fiable	*n* = 25	*n* = 50	*n* = 100
1. $\|TEI\| \geq .20$	75.27(9.44)	75.34(6.25)	75.15(4.58)
$\|TEI\| \geq .50$	43.82(11.43)	43.07(8.08)	43.13(5.82)
$\|TEI\| \geq .80$	21.95(9.73)	20.77(7.07)	20.85(5.03)
2. $\|DIT\| \geq .20$	84.49(6.95)	84.79(4.75)	84.24(3.39)
$\|DIT\| \geq .50$	62.53(8.59)	62.37(5.73)	61.84(4.27)
$\|DIT\| \geq .80$	43.30(7.60)	43.04(5.25)	42.77(3.62)
$\|DIT\| \geq 1.64$	10.28(3.66)	10.21(2.64)	10.03(1.78)
$\|DIT\| \geq 1.96$	4.86(3.06)	4.80(2.05)	4.92(1.46)
3. $\|DTW\| \geq 1$	61.96(10.20)	61.83(6.69)	61.50(5.05)
$\|DTW\| \geq 1.96$	33.02(9.55)	32.67(6.93)	32.80(4.94)
$\|DTW\| \geq 2.77$	16.83(7.82)	16.58(5.48)	16.61(3.97)
4. $\|ICF\| \geq 1.64$	24.61(9.34)	24.56(6.68)	24.65(4.78)
$\|ICF\| \geq 1.96$	16.73(8.13)	16.58(5.82)	16.50(4.25)
5. $\|GLN\| \geq 1.96$	16.42(7.54)	16.00(5.38)	16.06(3.88)
6. $\|HA\| \geq 1.65$*	5.31(6.53)	5.08(5.10)	4.90(3.78)
7. $\|EN\| \geq 1.96$	30.13(9.48)	29.79(6.65)	29.87(4.78)
8. $Y \leq ICPP_{inf}$, $Y \geq ICPP_{sup}$ (95%)	59.60(8.32)	71.80(5.81)	78.75(3.82)
9. $Y \leq ICPI_{inf}$, $Y \geq ICPI_{sup}$ (95%)	2.15(1.53)	3.95(1.71)	4.34(1.35)

*Debido al criterio $R_{DD} > .4$, el estadístico *HA* descarta del análisis el 32.5%, el 24.4% y el 15.5% de las muestras para tamaños muestrales de 25, 50 y 100 respectivamente.

Conviene señalar que la *distribución de los cambios*, es decir, la distribución de las diferencias pre-post, es una distribución normal (resultado de restar dos puntuaciones procedentes de poblaciones normales) con una media de aproximadamente cero y una desviación típica de aproximadamente 9.4 (ver Tabla 2). En este escenario, lo que cabe esperar es que el 95% de las puntuaciones (es decir, el 95% de las diferencias o cambios) se encuentre entre ±1.96 veces la desviación típica de las diferencias; y el 90%, entre ±1.64 veces la desviación típica de las diferencias. Este criterio es el que, en principio, debería servir para valorar la significación estadística de un cambio individual. Y, aunque no hemos encontrado que este criterio se utilice de forma habitual para valorar la significación estadística del cambio individual, es, de hecho, el criterio que mejor funciona (sin excluir del análisis ninguna muestra): cuando la diferencia individual tipificada (*DIT*, método número 2) se compara con los puntos críticos

±1.96 y ±1.64 se obtienen tasas de falsos positivos de, aproximadamente, el 5% y el 10%, que son justamente las tasas esperadas de falsos positivos cuando se aplican esos criterios (ver, en la Tabla 4, los resultados correspondientes al método número 2).

Frente a este criterio, los resultados obtenidos con el resto de métodos son, en principio, bastante llamativos: donde cabía esperar tasas de falsos positivos en torno al 5% o al 10% (dependiendo del nivel de confianza elegido), hemos encontrado tasas que, en la mayoría de los casos, están por encima del 20%, llegando (y sobrepasando) en algunos casos al 80%. Únicamente el método Hageman y Arrindell (*HA*) y el intervalo de confianza para el pronóstico individual de la regresión lineal (*ICPI*) ofrecen resultados que podríamos denominar aceptables; no obstante, el método *HA* lo hace a costa de descartar un importante número de muestras del análisis (>15.5%) y el *ICPI* solamente alcanza niveles aceptables de falsos positivos con $n = 50$ y $n = 100$, aunque es, junto con el estadístico *DIT*, el método que ofrece un mejor funcionamiento global.

Los resultados obtenidos con el *tamaño del efecto individual* (*TEI*) no deberían sorprender: se trata de un estadístico que pretende valorar el cambio individual con una estrategia diseñada para valorar diferencias entre promedios. Y ya sabemos que las puntuaciones individuales varían sensiblemente más que sus promedios. Para valorar correctamente la significación de los cambios individuales habría que aplicar criterios distintos de los utilizados para valorar el cambio en los promedios. De hecho, a una diferencia tipificada entre promedios de .20 puntos (punto de corte propuesto por Cohen y ampliamente aceptado como valor a partir del cual se pasa del no-efecto al efecto) le corresponde una diferencia individual tipificada (*DIT*) de .32 puntos, es decir, de .32 desviaciones típicas de las diferencias. Y no parece razonable asumir que una variación de .32 desviaciones típicas está reflejando un cambio estadísticamente significativo. Para obtener un resultado aceptable en términos de falsos positivos habría que valorar los cambios individuales pre-post tomando como criterio 1.2 desviaciones típicas, lo cual se corresponde con un efecto de tamaño grande (ver Hopkins 2002). Cuando el *TEI* se valora con el criterio de 1.2 desviaciones típicas del pre-tratamiento se obtiene, efectivamente, una tasa de falsos positivos que se ajusta a la esperada (en torno al 5% con niveles de confianza del 95%).

La *diferencia individual tipificada* (*DIT*) ofrece resultados igualmente inaceptables, pero sólo si se utilizan los criterios definidos por Cohen para valorar diferencias entre promedios. Ya hemos señalado que, al aplicar criterios convencionales (puntos críticos correspondientes a niveles de confianza del 90% y del 95%), la *DIT* ofrece los mejores resultados de todos los estadísticos evaluados.

Los métodos basados en el error de medida también ofrecen tasas inaceptables de falsos positivos. Por ejemplo, la *diferencia tipificada de Wirwich* (*DTW*) ofrece en torno a un 62% de falsos positivos con el criterio ±1, en torno a un 33% con el criterio ±1.96 y en torno a un 17% con el criterio ±2.77. Con los métodos *ICF*, *GLN* y *EN* se obtienen resultados parecidos. El *índice de cambio fiable* (*ICF*) de Jacobson y sus colaboradores, a pesar de ser el método más utilizado para valorar la presencia de un cambio significativo (ver Ogles et al., 2001), ofrece tasas de falsos positivos en torno al 25% donde cabría esperar un 10% y en torno a un 17% donde cabría esperar un 5% (aunque es más conservador que el método *DTW*, pues la forma de estimar el error típico de medida en el estadístico *ICF* siempre arroja un valor mayor que en la solución propuesta por Wyrwich). Incorporar el efecto de regresión a la media, que es lo que hace el método *GLN*, parece que no ayuda a corregir el problema (Brooks, Strauss, Sherman, Iverson y Slick, 2009, han argumentado que el efecto de regresión a la media es poco relevante en estos contextos). Y la solución basada en el intervalo de confianza de las puntuaciones verdaderas (*EN*) tampoco parece ser una estrategia adecuada: presenta más de un 30% de falsos positivos donde cabría esperar un 5%.

La complicación que añade el método *HA* lo convierte en el más conservador de los cinco basados en el error típico de medida. Al incorporar la fiabilidad de las diferencias se obtienen tasas de falsos positivos próximas al 5%, pero este resultado es engañoso, ya que para obtener esa tasa es necesario descartar las muestras en las que la fiabilidad de las diferencias (R_{DD}) es menor de .40 (criterio establecido por Hageman y Arrindell, 1999, p. 1173). En nuestro estudio ha sido necesario descartar el 32.5% de las muestras de tamaño 25, el 25.4% de las de tamaño 50 y el 15.5% de las de tamaño 100). Esta limitación es ya lo bastante importante en las condiciones concretas simuladas ($R_{XX} = .90$; y $R_{XY} = .80$), pero, dadas las características de R_{DD}, la pérdida de muestras sería todavía mayor si los valores de R_{XX} y R_{XY} fueran más parecidos.

Las altas tasas de falsos positivos obtenidas con los métodos basados en el error de medida no son del todo sorprendentes. Debe tenerse en cuenta que en la variación aleatoria no solo hay error de medida: los sujetos pueden experimentar pequeños cambios entre los momentos pre y post, y esos cambios, aunque no sean importantes ni clínica ni estadísticamente, sumados al error de medida pueden generar puntuaciones cuya variación aleatoria supere ampliamente dos errores típicos de medida. Esto es lo que parece reflejar el hecho de que todos los estadísticos basados en el error de medida presenten altas tasas de falsos positivos.

Una posible solución a este problema pasaría por estimar el error de medida a partir, no en la consistencia interna (alfa de Cronbach), sino de la estabilidad temporal (la correlación test-retest). Puesto que la consistencia interna tiende a ser mayor que la estabilidad temporal, estimar el error de medida a partir de la estabilidad temporal siempre arroja un error de medida más grande; y con un error de medida más grande se obtienen menos falsos positivos. Aunque la estrategia más recomendada y utilizada consiste en calcular el error de medida a partir de la consistencia interna (ver, por ejemplo, Bauer, Lambert y Nielsen, 2004; Martinovich, Sanunders y Howard, 1996; Tingey, Lambert, Burlingame y

Hansen, 1996), solo la estabilidad temporal estaría reflejando tanto las variaciones aleatorias del error de medida como las variaciones aleatorias atribuibles al paso del tiempo y a otras fuentes de variabilidad distintas del tratamiento. De hecho, cuando el error de medida se estima a partir de la correlación test-retest, en nuestros datos simulados se obtiene un valor parecido al error típico de medida de las diferencias individuales tipificadas; y, con ello, una tasa de falsos positivos más ajustada a lo esperable. A pesar de que Martinovich et al. (1996) recomiendan no utilizar la fiabilidad test-retest para calcular el error de medida en muestras clínicas, nuestros resultados sugieren que esa forma de calcular la fiabilidad es precisamente la que permite obtener tasas aceptables de falsos positivos. Habrá que recoger nueva evidencia sobre esta posibilidad.

Por último, los intervalos de confianza de los pronósticos de la regresión lineal funcionan mejor, tal como cabía esperar, cuando el pronóstico se considera una puntuación individual (*ICPI*) que cuando se considera una puntuación promedio (*ICPP*). Con el *ICPP*, la tasa de falsos positivos es del todo inaceptable; con el *ICPI* es razonablemente buena (mejora conforme va aumentando el tamaño muestral). De hecho, el *ICPI* es, junto con la *DIT*, el estadístico que mejor funciona. No obstante, tanto el *ICPP* como el *ICPI* tienen una propiedad poco deseable cuando se utilizan para evaluar el cambio individual: el valor de los límites de confianza (y, consiguientemente, la tasa de falsos positivos) depende del tamaño muestral.

Conclusión

El objetivo de este estudio era evaluar el comportamiento de varios métodos estadísticos (a los que hemos llamado *internos*) diseñados para estimar, a partir de las respuestas de los individuos, cuándo se produce un *cambio estadísticamente fiable*. Aunque para valorar un *cambio clínicamente significativo* parece claro que es necesario acompañar estos métodos internos de algún criterio o referente externo (Crosby et al., 2003; de Vet et al., 2007; Jacobson y Truax, 1991; Kolotkin et al., 2002; Sheldrick et al., 2001), también parece claro que los métodos internos deberían actuar como filtro de los cambios que únicamente representan fluctuaciones atribuibles al azar muestral.

El problema es que los resultados obtenidos dejan en mal lugar a la mayoría de los métodos internos evaluados: la tasa observada de falsos positivos alcanza valores inaceptables con todos ellos, exceptuando los métodos *HA* e *ICPI*, de cuyas limitaciones ya hemos hablado, y el método *DIT*, que únicamente funciona bien cuando se aplican criterios que no son los habitualmente utilizados. Esto significa que, entre los métodos habitualmente utilizados para valorar la significación estadística del cambio individual (es decir, entre los índices de cambio basados en el error típico de medida), ninguno de ellos consigue el objetivo de identificar y desechar las variaciones propias del azar muestral. Y esto implica que, al utilizar estos métodos para valorar cuándo se produce un cambio estadísticamente fiable, se están considerando estadísticamente fiables cambios que únicamente responden a variaciones aleatorias.

El problema es más grave si se tiene en cuenta que algunos métodos de los que hemos llamado *externos* (métodos que, combinados con los internos, se utilizan para decidir cuándo se ha producido un cambio clínicamente significativo) pueden llevar a tantos o más falsos positivos que los métodos internos evaluados en este estudio. Es decir, muchos de los métodos externos que se utilizan en combinación con los internos, no sirven para identificar y descartar los falsos positivos que los métodos internos pasan por alto (Turner et al., 2010).

Estos resultados son todavía más preocupantes cuando se considera que en los últimos años ha tomado fuerza (particularmente en contextos médicos) la idea de que el cambio clínicamente significativo (CCS) o la diferencia mínimamente importante (DMI) se corresponde con $ETM = 1$ o con $TEI = .5$ (Turner et al., 2010). Varios estudios han encontrado coincidencias entre la estimación del CCS o de la DMI con dichos criterios (Norman, Sloan y Wyrwich, 2003; Rejas et al., 2008; Wyrwich, 2004; Wyrwich et al., 1999). Esta equivalencia entre las valoraciones de los métodos internos y externos (recordemos que, en el presente estudio, los métodos internos han presentado tasas excesivamente altas de falsos positivos) podría estar evidenciando la presencia de sesgos en los métodos externos*:* los pacientes suelen presentar sesgos mnémicos que favorecen las percepciones positivas sobre los efectos de un tratamiento (ver, por ejemplo, Cella et al., 2002).

Nuestros resultados indican que, para decidir que se ha producido un *cambio estadísticamente fiable*, la mejor estrategia consiste en aplicar el criterio 1.96(*DIT*) o, lo que resulta equivalente cuando las desviaciones típicas del pre y del post son similares y la correlación test-retest es aproximadamente .8, el criterio 1.2(*TEI*). De los restantes métodos evaluados, únicamente el *ICPI* ofrece resultados que se aproximan a los obtenidos con la *DIT*.

No obstante, esta recomendación debe tomarse con cautela hasta obtener suficiente evidencia sobre dos aspectos importantes no contemplados en este estudio y que necesitan ser investigados. En primer lugar es necesario valorar el funcionamiento de estos métodos en condiciones de no-normalidad y utilizando diferentes niveles de consistencia interna y fiabilidad test-retest; esto permitiría disponer de resultados más generalizables. En segundo lugar, es necesario valorar el funcionamiento de todos estos métodos en situaciones de cambio para poder identificar con precisión la tasa de falsos negativos asociada a cada método (pues es importante no pasar por alto el hecho de que la utilización de criterios conservadores para controlar correctamente la tasa de falsos positivos podría llevar a obtener tasas elevadas de falsos negativos y, consecuentemente, a no detectar cambios clínicamente relevantes que de hecho se estarían produciendo).

Agradecimiento.- Este trabajo ha sido posible gracias al Convenio de Desempeño UTA-MINEDUC.

Referencias

Abad, F., Olea, J., Ponsoda, V. y García, C. (2011). *Medición en ciencias sociales y de la salud*. Madrid: Síntesis.

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: LEA.

Angst, F., Verra, M. L., Lehmann, S. y Aeschlimann, A. (2008). Responsiveness of five condition-specific and generic outcome assessment instruments for chronic pain. *Medical Research Methodology*, *8*, 26 (8).

Atkins, D. C., Bedics, J. D., McGlinchey, J. B. y Beauchaine, Th. P. (2005). Assessing clinical significance: Does it matter which method we use? *Journal of Consulting and Clinical Psychology*, *73*, 982-989.

Barlow, D. H. (1981). On the relation of clinical research to clinical practice: Current issues, new directions. *Journal of Consulting and Clinical Psychology*, *49*, 147-155.

Bauer, S., Lambert, M. J. y Nielsen, S. L. (2004). Clinical significance methods: A comparison of statistical techniques, *Journal of Personality Assessment*, *82*, 60-70.

Bergin, A. E. (1971). The evaluation of therapeutic outcomes. En A. E. Bergin y S. L. Garfield (Eds.), *The handbook of psychotherapy and behavior change*. New York: Wiley.

Bergin, A. E. y Lambert, M. J. (1978). The evaluation of therapeutic outcomes. En S.L. Garfield y A.E. Bergin (Eds.), *The handbook of psychotherapy and behavior change* (2ª ed.). New York: John Wiley.

Brooks, B. L., Strauss, E., Sherman, E. M., Iverson, G. L. y Slick, D. J. (2009). Developments in neuropsychological assessment: Refining psychometric and clinical interpretive methods. *Canadian Psychology*, *50*, 196-209.

Cella, D., Hahn, E. A., y Dineen, K. (2002). Meaningful change in cancer-specific quality of life scores: Differences between improvement and worsening. *Quality of Life Research*, 11, 207-221.

Childs, J. D., Piva, S. R. y Fritz, J. M. (2005). Responsiveness of the numeric pain rating scale in patients with low back pain. *Spine*, *30*, 1331-1334.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2ª ed.). New York: Academic Press.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.

Crawford, J. R. y Howell, D. C. (1998). Regression equations in clinical neuropsychology: An evaluation of statistical methods for comparing predicted and obtained scores. *Journal of Clinical and Experimental Neuropsychology*, *20*, 755-762.

Crawford, J. R. y Garthwaite, P. H. (2006). Comparing patients' predicted test scores from a regression equation with their obtained scores: A significance test and point estimate of abnormality with accompanying confidence limits. *Neuropsychology, 20*, 259-271.

Crosby, R. D., Kolotkin, R. L. y Williams, G. R. (2003). Defining clinically meaningful change in health-related quality of life. *Journal of Clinical Epidemiology*, *56*, 395-407.

Crosby, R. D., Kolotkin, R. L. y Williams, G. R. (2004). An integrated method to determine meaningful changes in health-related quality of life, *Journal of Clinical Epidemiology*, *57*, 1153-1160.

de Vet, H. C., Terluin, B., Knol, D. L., Roorda, L. D., Mokkink, L. B., Ostelo, R. W., Hendriks, E. J., Bputer, L.M. y Terwee C. B. (2010). Three ways to quantify uncertainty in individually applied "minimally important change" values. *Journal of Clinical Epidemiology*, *63*, 37-45.

de Vet, H. C., Terwee, C. B., Ostelo, R. W., Beckerman, H., Knol, D. L. y Bouter, L. M. (2006). Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health and Quality of Life Outcomes*, *4*, 54(5).

de Vet, H. C., Ostelo, R. W., Terwee, C. B., van der Roer, N., Knol, D. L., Beckerman, H., Boers, M. y Bouter, L. M. (2007). Minimally important change determined by a visual method integrating an anchor-based and a distribution-based approach. *Quality of Life Research*, *16*, 131-142.

Edwards, D. W., Yarvis, R. M., Mueller, D. P., Zingale, H. C. y Wagman, W. J. (1978). Test-taking and the stability of adjustment scales: Can we assess patient deterioration? *Evaluation Quarterly*, *2*, 275-292.

Eton, D. T., Cella, D., Yost, K. J., Yount, S. E., Peterman, A. H., Neuberg, D. S., Sledge, G. W. y Wood, W. C. (2004). A combination of distribution- and anchor-based approaches determined minimally important differences (MIDs) for four endpoints in a breast cancer scale. *Journal of Clinical Epidemiology*, *57*, 898-910.

Fayers, P. M. y Machin, D. (2000). *Quality of life: Assessment, analysis and interpretation*. Chichester: Wiley.

Gatchel, R. J. y Mayer, T. G. (2010). Testing minimal clinically important difference: consensus or conundrum? *Spine Journal*, *10*, 321-327.

Guyatt, G. H., Bombardier, C., Tugwell, P. X. (1986). Measuring disease-specific quality of life in clinical trials. *Canadian Medical Association Journal*, *134*, 889-895.

Guyatt, G. H., Osoba, D., Wu, A. W., Wyrwich, K. W., Norman, G. R. y el grupo Clinical Significance Consensus Meeting (2002). Methods to explain the clinical significance of health status measures. *Mayo Clinic Proceedings*, *77*, 371-383.

Hageman, W. J. y Arrindell, W. A. (1999). Establishing clinically significant change: Increment of precision and the distinction between individual and group level of analysis. *Behavior Research and Therapy*, *37*, 1169-1193.

Hopkins W. G. (2002). *A scale of magnitudes for effect statistics*. Disponible en Society for Sports Science website: http://www.sportsci.org/resource/stats/effectmag.html.

Hsu, L. M. (1989). Reliable changes in psychotherapy: Taking into account regression toward the mean. *Behavioral Assessment*, *11*, 459-467.

Hsu, L. M. (1995). Regression toward the mean associated with measurement error and the identification of improvement and deterioration in psychotherapy. *Journal of Consulting and Clinical Psychology*, *63*, 141-144.

Hsu, L. M. (1996). On the identification of clinically significant client changes: Reinterpretation of Jacobson's cut scores. *Journal of Psychopathology and Behavioral Assessment*, *18*, 371-385.

Hsu, L. M. (1999). Caveats concerning comparisons of change rates obtained with five methods of identifying significant client changes: Comment on Speer and Greenbaum (1995). *Journal of Consulting and Clinical Psychology*, *67*, 594-598.

Jacobson, N. S., Follette, W. C. y Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, *15*, 336-352.

Jacobson, N. S., Follette, W. C. y Revenstorf, D. (1986). Toward a standard definition of clinically significant change. *Behavior Therapy*, *17*, 308-311.

Jacobson, N. S., Roberts, L. J., Berns, S. B. y McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology*, *67*, 300-307.

Jacobson, N. S. y Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12-19.

Jaeschke, R., Singer, J. y Guyatt, G. H. (1989) Measurement of health status: ascertaining the minimal clinically important difference. *Controlled Clinical Trials*, *10*, 407-415.

Kazdin, A. E. (1977). Assessing the clinical or applied importance of behavior change through social validation. *Behavior Modification*, *1*, 427–452.

Kazdin, A. E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting and Clinical Psychology*, *67*, 332-339.

Kazdin, A. E. (2001). Almost clinically significant ($p <.10$): Current measures may only approach clinical significance. *Clinical Psychology: Science and Practice*, *8*, 455-462.

Kendall, P. C. (1997). Editorial. *Journal of Consulting and Clinical Psychology*, *65*, 3-5.

Kendall, P. C. (1999). Clinical significance. *Journal of Consulting and Clinical Psychology*, *67*, 283-284.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746-759.

Kolotkin, R. L., Crosby, R. D., Williams, G. R. (2002). Integrating anchor-based and distribution-based methods to determine clinically meaningful change in obesity-specific quality of life. *Quality of Life Research*, *11*, 670.

Lambert, M. J. y Ogles, B. M. (2004). The efficacy and effectiveness of psychotherapy. En M. J. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (pp. 139-193). New York: Wiley.

Lydick, E. y Epstein, R. (1993) Interpretation of quality of life changes. *Quality of Life Research*, *2*, 221-226.

Martinovich, Z., Saunders, S. y Howard, K. (1996). Some comments on "Assessing clinical significance". *Psychotherapy Research*, *6*, 124-132.

Mathias, S. D., Crosby, R. D., Qian, Y., Jiang, Q., Dansey, R. y Chung, K. (2011). Estimating minimally important differences for the worst pain rating of the Brief Pain Inventory-Short Form. *The Journal of Supportive Oncology*, *9*, 72-78

McGlinchey, J. B., Atkins, D. C. y Jacobson, N. S. (2002). Clinical significance methods: which one to use and how useful are they? *Behavior Therapy*, *33*, 529-550.

Norman, G. R., Sloan, J. A. y Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Medical Care*, *41*, 582-592.

Norman, G. R., Sridhar, F. G., Guyatt, G. H. y Walter, S.D. (2001). Relation of distribution- and anchor-based approaches in interpretation of changes in health-related quality of life. *Medical Care*, *39*, 1039-1047.

Norman, G. R., Stratford, P. y Regehr, G. (1997). Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. *Journal of Clinical Epidemiology*, *50*, 869-879.

Nunnally, J. C. y Kotsch, W. E. (1983). Studies of individual subjects: Logic and methods of analysis. *British Journal of Clinical Psychology*, *22*, 83-93.

Nunnally, J. (1975). The study of change in evaluation research: Principles concerning measurement, experimental design and analysis. En E. L. Streuning y M. Guttentag (Eds.), *Handbook of Evaluation Research*. Beverly Hills, CA: Sage.

Ogles, B. M., Lunnen, K. M. y Bonesteel, K. (2001). Clinical significance: History, application, and current practice. *Clinical Psychology Review*, *21*, 421-446.

Rejas, J., Pardo, A. y Ruiz, M. (2008). Standard error of measurement as a valid alternative to minimally important difference for evaluating the magnitude of changes in patient-reported outcomes measures. *Journal of Clinical Epidemiology*, *61*, 350-356.

Revicki, D., Hays, R. D., Cella D. y Sloan, J.A. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology*, *61*, 102-109.

Sheldrick, R. C., Kendall, P. C. y Heimberg, R. G. (2001). The clinical significance of treatments: A comparison of three treatments for conduct disordered children. *Clinical Psychology: Science and Practice*, *8*, 418-430.

Speer, D. C. (1992). Clinically significant change: Jacobson and Truax (1991) revisited. *Journal of Consulting and Clinical Psychology*, *60*, 402-408.

Speer, D. C. y Greenbaum, P. H. (1995). Five methods for computing significant individual client change and improvement rates: Support for an individual growth curve approach. *Journal of Consulting and Clinical Psychology*, *63*, 1044-1048.

Stucki, G., Liang, M., Fossel, A. y Katz, J. (1995). Relative responsiveness of condition specific and generic health status measures in degenerative lumbar spinal stenosis. *Journal of Clinical Epidemiology*, *48*, 1369-1378.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education*, *61*, 361-377.

Thompson, B. (2002). 'Statistical', 'practical' and 'clinical': How many kinds of significance do counselors need to consider. *Journal of Counseling and Development*, *80*, 64-71.

Tingey, R., Lambert, M. J., Burlingame, G. y Hansen, N. (1996). Assessing clinical significance: Proposed extensions to method. *Psychotherapy Research*, *6*, 109-123.

Turner, D., Schünemann, H., Griffith, L., Beaton, D., Griffiths, A., Critch, J. y Guyatt, G. (2010). The minimal detectable change cannot reliably replace the minimal important difference. *Journal of Clinical Epidemiology*, *63*, 28-36.

Williams, P., Weiss, L. y Rolfhus, E. (2003). *WISC-IV. Technical report nº 2: Psychometric properties*. San Antonio, TX: The Psychological Corporation.

Wyrwich, K. (2004). Minimal important difference thresholds and the standard error of measurement: Is there a connection? *Journal of Biopharmaceutical Statistics*, *14*, 97-110.

Wyrwich, K. W., Tierney, W. M. y Wolinsky, F. D. (1999). Further evidence supporting an SEM-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *Journal of Clinical Epidemiology*, *52*, 861-873.

Wyrwich, K. W. y Wolinsky, F. D. (2000). Identifying meaningful intra-individual change standards for health-related quality of life measures. *Journal of Evaluation in Clinical Practice*, *6*, 39-49

Yost, K. J., Cella, D., Chawla, A., Holmgren, E., Eton, D. T., Ayanian, J. Z. y West, D. W. (2005). Minimally important differences were estimated for the Functional Assessment of Cancer Therapy-Colorectal (FACT-C) instrument using a combination of distribution- and anchor-based approaches. *Journal of Clinical Epidemiology*, 58, 1241-1251.

(Artículo recibido: 07-11-2011, revisado: 18-06-2012, aceptado: 18-06-2012)